DOUBLE SAMPLING
AND
THE CURTIS IMPACT STUDY

by

D.S. Robson and A.J. King

BU-13-M Dec. 1950

INTRODUCTION

The expression "double sampling" has acquired a variety of meanings in modern statistics and it is necessary before beginning any discussion on the subject to describe in some detail the sampling technique to be considered. One is interested in estimating the population mean of the variable Y; a preliminary sample is drawn on a related variable X and the information thus obtained is used either as a guide to the selection of a smaller sample on the variable Y or as a means of adjusting the estimate computed from a smaller sample on Y. In accordance with the historical development of the theory, double sampling may be roughly classified into three uses:

Use I - stratification, i.e., the large sample is stratified on the basis of the variable X and then subsampled for measurement of Y

Use II - regression estimation, i.e., the mean of the large sample on X is used to adjust the mean of the smaller sample on Y

Use III - matched sampling, i.e., the same population is sampled at two points in time and the second sample consists partly of elements common to the first sample.

In general, measurement of the variable Y is more costly than measurement of X.

Three particular problems relating to double sampling shall be treated in this discussion. The first is optimum allocation of resources to the large and small samples under Use II; the second relates to double sampling from equal sized clusters and the corresponding regression estimate; and the third is an application to a binomial type population which was sampled in the 1948 Curtis Impact Study.

REVIEW OF THE LITERATURE ON DOUBLE SAMPLING

Use I

Neyman (1938) first described this technique in an approach to the case where the variable Y is difficult to measure while information on the related variable X is easily obtained or already available. The procedure is outlined as follows:

(i) a relatively large random sample of size N is drawn and measured for the variable X,

(ii) these N elements are divided into $\underline{k}$ groups of size $N_1, N_2, \ldots N_k$ such that the elements within each group are as homogeneous as is practicable with respect to the variable X,

(iii) random samples of size $n_1, n_2, \ldots, n_k$ are then drawn from each group and measured for the variable Y, the within group sample size, n_i, being taken proportional to the size of the group, i.e.,

(1) $$n_i = n\frac{N_i}{N} = w_i n ,$$

where w_i is then an unbiased estimate of the true proportion in the population, $W_i = \frac{M_i}{M}$.

The population mean for the variable Y is given by

(2) $$\mu_y = \sum_{i=1}^{k} W_i \mu_{yi}$$

and an unbiased estimate of μ_y is

(3) $$\bar{y}_p = \sum_{i=1}^{k} w_i \bar{y}_{ni} .$$

If the finite population corrections (f.p.c.'s) are ignored, the variance of this estimate may be written in the form

(4) $$V(\bar{y}_p) = \sum_{i=1}^{k} \left[\frac{\sigma_i^2}{n_i} \left\{ W_i^2 + \frac{W_i(1-W_i)}{N} \right\} + \frac{W_i(\mu_{yi}-\mu_y)^2}{N} \right] \quad *$$

where σ_i^2 is the true variance of Y in the i'th group.

If the variance of this estimate is to be compared with the variance of an estimate obtained from a single random sample in which Y alone is measured then cost considerations must be brought into account; clearly, for a fixed cost the size of the single sample on Y could be considerably different from the size of the stratified sample. Neyman therefore set up a simple cost function of the form

(5) $$C = nA + NB ,$$

where C = total cost of the sample

A = fixed cost of measuring the variable Y on a single sample element

B = fixed cost of measuring the variable X on a single sample element.

He then showed that $V(\bar{y}_p)$ is approximately minimized for

(6) $$\frac{n}{N} = \frac{\sum W_i \sigma_i}{\sqrt{\frac{B}{A \sum W_i(\mu_{yi}-\mu_y)^2}}} \quad **$$

Use II

Cochran [Snedecor and King (1941), Cochran (1941)] and Bose (1941, 1943) were the first to consider double sampling in relation to regression estimation. The theory which they developed is based upon the linear model

(7) $$Y = \mu_y + \beta(X-\mu_x) + \varepsilon$$

---

* A development of this equation is given in Appendix I

** A development of this approximation is given in Appendix II

where μ_y = the population mean of the variable Y

β = the population regression coefficient

μ_x = the population mean of the variable X

ε = the error term with mean zero and variance σ_ε^2 .

The parameter of interest is μ_y, and its estimate, denoted by $\bar{y}_p$, is of the form

$$(8) \qquad \bar{y}_p = \bar{y}_n + b(\bar{x}_N - \bar{x}_n)$$

where $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$

$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\bar{x}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$

$b = \dfrac{\sum_{1}^{n} (X_i - \bar{x}_n)(Y_i - \bar{y}_n)}{\sum_{1}^{n} (X_i - \bar{x}_n)^2}$.

In the special case when the X's are considered fixed the variance of the estimate is

$$(9) \qquad V(\bar{y}_p) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_{1}^{n} (X_i - \bar{x}_n)^2} \right] + \beta^2 (\bar{x}_N - \mu_x)^2 \ * \ .$$

Since this variance depends upon the particular sets of X chosen in the sample it is not comparable with the variances which would arise from other sampling procedures. In order to determine the variance of $\bar{y}_p$ which would be comparable with, for example, the variance of the mean of a simple random sample on the variable Y, it is necessary to find the average value of $V(\bar{y}_p)$ when the X's

---

* A development of (9) is given in Appendix III

are allowed to vary over the range in the entire population. Under the assumptions that (1) the large sample of size N is drawn at random, (2) the small sample of size n is drawn at random from the large sample, and (3) the variable X is normally distributed, the expected value of $V(\bar{y}_p)$ is

$$(10) \qquad V(\bar{y}_p) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \left(\frac{1}{n} - \frac{1}{N} \right)\left(\frac{1}{n-3} \right) \right] + \frac{\beta^2 \sigma_x^2}{N} \quad * \quad .$$

If assumption (2) is altered to read: the small sample is drawn at random and independently of the large sample, then

$$(11) \qquad V(\bar{y}_p) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \left(\frac{1}{n} + \frac{1}{N} \right)\left(\frac{1}{n-3} \right) \right] + \frac{\beta^2 \sigma_x^2}{N}$$

as given by Bose (1943).

Schumacher (1942, 1948) discusses the case mentioned earlier in which the two samples of size n and N are drawn independently. The estimate (8) is given, i.e.,

$$\bar{y}_p = \bar{y}_n + b(\bar{x}_N - \bar{x}_n) \; ;$$

however, the variance of this estimate is stated to be

$$(12) \qquad V(\bar{y}_p) = \frac{\sigma_\varepsilon^2}{n} + b^2 \frac{\sigma_x^2}{N} + (\bar{x}_N - \bar{x}_n)^2 \sigma_b^2$$

$$\text{where } \sigma_b^2 = \frac{\sigma_\varepsilon^2}{\sum_1^n (x_i - \bar{x}_n)^2} \quad ,$$

$$(13) \qquad \text{or} \quad V(\bar{y}_p) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_1^n (x_i - \bar{x}_n)^2} \right] + \frac{b^2 \sigma_x^2}{N} \quad .$$

This latter expression is a composite of sample values and parameters; i.e., it contains the statistic b with the parameter σ_b^2 and the statistic $(\bar{x}_N - \bar{x}_n)$

---

* A development of (10) is given in Appendix IV.

with the parameter σ_x^2. As such, its nearest counterpart in the literature is a "hybrid" formula given by Cochran (1948):

$$(14) \qquad V(\bar{y}_p) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_1^n (X_i - \bar{x}_n)^2} \right] + \frac{\beta^2 \sigma_x^2}{N}$$

$$(15) \qquad = \sigma_y^2 (1 - \rho^2) \left[\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_1^n (X_i - \bar{x}_n)^2} \right] + \frac{\rho^2 \sigma_y^2}{N} \quad .$$

Schumacher suggests dropping the term $\dfrac{\sigma_\varepsilon^2 (\bar{x}_N - \bar{x}_n)^2}{\sum_1^n (X_i - \bar{x}_n)^2}$ from (13) in deriving an approximate solution to the optimum value of the ratio $\frac{n}{N}$. The cost function is given as before,

$$C = nA + NB \ ,$$

and the solution is given as

$$(16) \qquad \frac{n}{N} = \frac{\sigma_\varepsilon}{b\sigma_x} \sqrt{\frac{B}{A}} \quad .$$

When b is replaced by β and subsequently $\beta^2 \sigma_x^2$ is replaced by $\rho^2 \sigma_y^2$, then (16) may be written as

$$(17) \qquad \frac{n}{N} = \sqrt{\frac{B(1 - \rho^2)}{A\rho^2}}$$

and in this form is identical to the approximate solution considered in section 2 of this paper.

Use III

Jessen (1942) also worked on the problem of optimum allocation in double sampling using Cochran's formulae

$$\bar{y}_p = \bar{y}_n + b(\bar{x}_N - \bar{x}_n)$$

$$V(\bar{y}_p) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \left(\frac{1}{n} - \frac{1}{N} \right)\left(\frac{1}{n-3} \right) \right] + \frac{\rho^2 \sigma_y^2}{N} \quad .$$

The particular application of double sampling described by Jessen is the technique known as matched sampling; this procedure attempts to measure and utilize changes in a population over a period of time. At some beginning point in time a large sample of size N is drawn and considered to be a sample on the variable X. Subsequently, a subsample of n out of these N elements is drawn and this is called the matched sample; measurement of the elements at this second point in time is considered to a measure on the variable Y. In addition, one may have an unmatched sample of size m on the variable Y; i.e., a sample drawn independently of the first large sample. The mean of the unmatched sample is denoted by

$$\bar{y}_u = \frac{1}{m} \sum_{j=1}^{m} Y_j \quad , \tag{18}$$

and the weighted estimate (weighted inversely to the variances) of the population mean at the second point in time is

$$\bar{y}_w = \frac{\bar{y}_p \, V(\bar{y}_u) + \bar{y}_u \, V(\bar{y}_p)}{V(\bar{y}_u) + V(\bar{y}_p)} \quad . \tag{19}$$

The variance of this weighted estimate is

$$V(\bar{y}_w) = \frac{V(\bar{y}_u) \, V(\bar{y}_p)}{V(\bar{y}_u) + V(\bar{y}_p)} \quad . \tag{20}$$

Jessen then attacked the problem of determining the sample sizes n and m which minimize $V(\bar{y}_w)$ under the conditions that m + n and N are fixed. The approximate solution which he arrived at was

$$\frac{n^2}{m^2} = 1 - \rho^2 \; * . \tag{21}$$

---

* An alternative to this approximation is given in Appendix V.

OPTIMUM ALLOCATION TO THE LARGE AND SMALL SAMPLES

(Use II)

The cost function (5) given by Neyman,

$$C = nA + NB,$$

and the estimate (8),

$$\bar{y}_p = \bar{y}_n + b(\bar{x}_N - \bar{x}_n) ,$$

with variance (10),

$$V(\bar{y}_p) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{1}{n-3}\right)\right] + \frac{\beta^2\sigma_x^2}{N}$$

given by Cochran lead directly to the solution for n and N which minimize the variance of the estimate subject to the fixed cost conditions. Following the customary procedure of introducing the Lagrangian multiplier, we have

(22) $$F(n,k,\lambda) = V(\bar{y}_p) + \lambda(C - nA - NB)$$

and minimization of (22) gives the equation

(23) $$\rho^2 = \frac{Bn^2[C - A(2n-3)] - (n^2-4n+6)(C - nA)^2}{Bn^2[C - A(2n-3)] - (n^2-4n+6)(C - nA)^2 - ABn^2(n-3)^2} .$$

Since this is a fourth degree equation in n, with rather awkward symbolical coefficients, the algebraic solution proved too difficult to warrant effort; the alternative was to find an approximate solution and then investigate the suitability of the approximation. Examination of (10) suggests that a reasonable estimate of the optimum value of n might be obtained by dropping the term $\sigma_\varepsilon^2 \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{1}{n-3}\right)$ from (22) and then differentiating. This procedure leads to an estimate of the form

(24) $$n \sim \frac{C}{\left(AB \frac{\rho^2}{1-\rho^2}\right)^{\frac{1}{2}} + A}$$

and later computations indicate that this is a very close approximation to the solution of (23).

Equation (23) describes the relationship among A, B, C, ρ, and n when the sampling error is at a minimum; hence, one may supply appropriate numerical values to the constants A, B, and C and some possible value of n to determine the value of ρ^2 under which conditions are optimum. The possible values of n include all integers in the closed interval $\left[4, \frac{C}{A+B}\right]$. The lower limit is set by the fact that $V(\bar{y}_p)$ is undefined at n = 3. The upper limit is set by the relation $n \leq N$ since the small sample is defined to be drawn from the large sample.

At a starting point in the investigation of the relationship between the correlation coefficient ρ and the optimum value of n under fixed cost conditions, the numerical values A = 2.5, B = 0.1, and C = 100 were arbitrarily assigned; n, then, might take on any of the values 4, 5, 6, ..., 38. When costs are thus fixed, each value of n determines the corresponding value of N and also the corresponding value of ρ at which the two sample sizes, n and N, would be optimum. Likewise, the value of $\frac{V(\bar{y}_p)}{\sigma_y^2}$ is determined when the quantities ρ, n, and N are given. The efficiency of the optimally allocated double sample relative to a single sample on Y may then be evaluated by comparing the variance of the two estimates. Under these numerical cost conditions the size of a single random sample on Y would be

$$n_r = \frac{C}{A} = \frac{100}{2.5} = 40,$$

and the estimate of the population mean, μ_y, would be

(25) $$\bar{y}_{n_r} = \frac{1}{n_r} \sum_{i=1}^{n_r} Y_i$$

with variance

(26) $$V(\bar{y}_{n_r}) = \frac{\sigma_y^2}{n_r} = \frac{\sigma_y^2}{40} = .025\ \sigma_y^2 \ .$$

The relative efficiency is therefore

(27) $$R.E. = 100 \left[\frac{V(\bar{y}_{n_r})/\sigma_y^2}{V(\bar{y}_p)/\sigma_y^2} \right] = \left[\frac{.025}{V(\bar{y}_p)/\sigma_y^2} \right] 100 \ .$$

Thus, $n = 10$ and $N = \frac{C - nA}{B} = \frac{100 - 10(2.5)}{0.1} = 750$ are the optimum sample sizes

if ρ^2 is [by equation (23)]

$$\rho^2 = \frac{0.1(10)^2\,[100-2.5(20-3)] - [(10)^2-4(10)+6]\,[100-2.5(10)]}{0.1(10)^2[100-2.5(20-3)] - [(10)^2-4(10)+6][100-2.5(10)] - 0.1(2.5)(10)^2(10-3)^2}$$

$\rho^2 = .9967061037$ $\qquad$ $(\rho = \pm .998351693)$

in which case $\frac{V(\bar{y}_p)}{\sigma_y^2}$ is [by equation (10)]

(28) $$\frac{V(\bar{y}_p)}{\sigma_y^2} = (1 - \rho^2) \left[\frac{1}{n} + \left(\frac{1}{n} - \frac{1}{N} \right)\left(\frac{1}{n-3} \right) \right] + \frac{\rho^2}{N}$$

$$= .0032938963 \left[\frac{1}{10} + \left(\frac{1}{10} - \frac{1}{750} \right)\left(\frac{1}{7} \right) \right] + \frac{.9967061037}{750}$$

$$= .0017047594$$

and the relative efficiency is [by equation (27)]

$$R.E. = 100 \left[\frac{.025}{.0017047594} \right] = 1466.48\ \%$$

The reliability of the approximation for n may now be investigated for this particular case in which $\rho^2 = .9967061037$. Equation (24) gives

$$n \sim \frac{100}{\left[\frac{0.1(2.5)(.9967061037)}{.0032938963}\right]^{\frac{1}{2}} + 2.5} = 8.930498$$

and the corresponding value of N is

$$N \sim \frac{100 - 8.930498(2.5)}{0.1} = 776.737543.$$

With this combination of ρ^2, n, and N, equation (28) gives

$$\frac{V(\bar{y}_p)}{\sigma_y^2} = .0017135105$$

and the loss in efficiency due to the use of this approximate solution for the optimum value of n is

$$100\% - 100\left[\frac{.0017047594}{.0017135105}\right] = 0.51\% \ .$$

This loss would, of course, have been even smaller were the integer 9 used in place of 8.930498.

Several tables and graphs have been prepared to show the behavior of the efficiency of the optimally allocated double sample relative to a single sample as ρ varies continuously in the interval $-1 \leq \rho \leq 1$. Table 1 and Figure 1 present the relationship between $|\rho|$ and the R. E. under the cost conditions already mentioned; i.e., C = ¢100, A = ¢2.50, B = ¢0.10.

TABLE 1

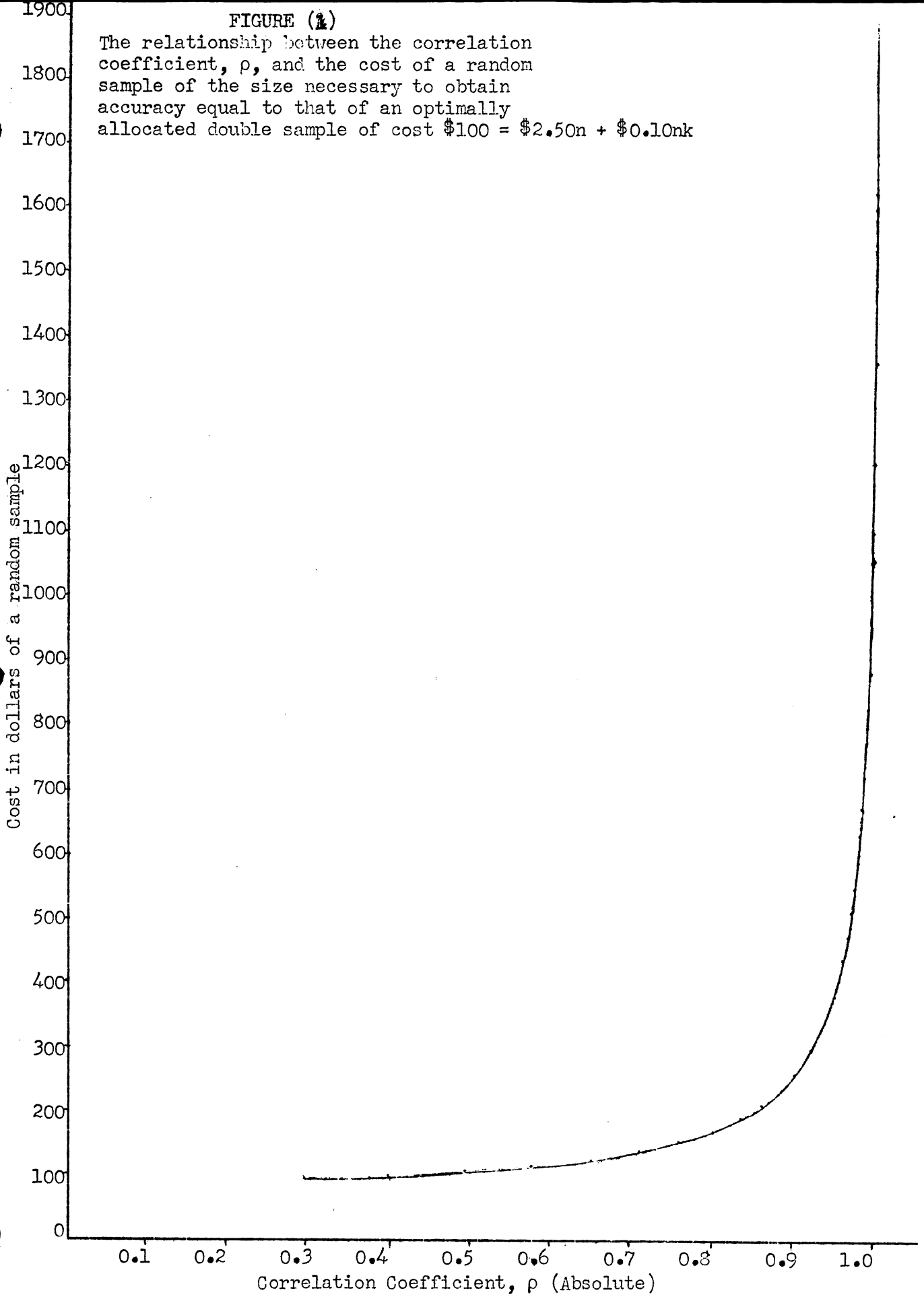
EFFICIENCY OF DOUBLE SAMPLING RELATIVE TO SINGLE SAMPLING

FOR C = 100 = 2.5n + 0.1N

$\lvert\rho\rvert$	Optimum n	Relative Efficiency	Approximation To Optimum n	Loss in Efficiency Due To Approximation
.999958731	4	2169.77%	1.738028	———
.999851240	5	2029.89	3.176125	———
.999688058	6	1902.43	4.441842	3.49%
.999466116	7	1784.28	5.619466	1.62
.999177184	8	1673.05	6.748607	0.99
.998810219	9	1567.38	7.848802	0.69
.998351693	10	1466.48	8.930498	0.51
.997785327	11	1369.90	9.999712	0.40
.997091573	12	1277.36	11.060118	0.32
.996246894	13	1188.65	12.114081	0.27
.995222849	14	1103.67	13.163181	0.22
.993984907	15	1022.31	14.207141	0.19
.992490952	16	944.52	15.250857	0.16
.990689343	17	870.25	16.290775	0.14
.988516436	18	799.46	17.328688	0.12
.985893348	19	732.14	18.364907	0.11
.982721760	20	668.25	19.399667	0.10
.978878409	21	607.79	20.433145	0.09
.974207863	22	550.73	21.465462	0.08
.968512990	23	497.07	22.496712	0.07
.961542408	24	446.81	23.526904	0.06
.952973972	25	399.92	24.556186	0.05
.942393211	26	356.41	25.584416	0.05
.929265583	27	316.28	26.611734	0.04
.912901719	28	279.51	27.637591	0.04
.892416044	29	246.11	28.662274	0.04
.866682131	30	216.07	29.685383	0.03
.834294779	31	189.39	30.706535	0.03
.793561399	32	166.06	31.725142	0.03
.742566564	33	146.09	32.740270	0.03
.679382793	34	129.48	33.750364	0.03
.602526817	35	116.21	34.752674	0.03
.511760476	36	106.30	35.741895	0.04
.4152	36.944	100.00	———	———

FIGURE (1)

The relationship between the correlation coefficient, ρ, and the cost of a random sample of the size necessary to obtain accuracy equal to that of an optimally allocated double sample of cost \$100 = \$2.50n + \$0.10nk



As one would expect, there exists a unique value of $|\rho|$, say $|\rho_c|$, such that R. E. $\leq 100\%$ for all $|\rho| \leq |\rho_c|$; that is, if the correlation coefficient for a bivariate normal population is numerically less than the critical value $|\rho_c|$ then double sampling can only be less efficient than single sampling. An algebraic expression of $|\rho_c|$ cannot be obtained explicitly without the aid of an exact solution for n in equation (23). It is obvious from the relationship

$$R.E. = 100\% = 100 \left[\frac{V(\bar{y}_{n_r})/\sigma_y^2}{V(\bar{y}_p)/\sigma_y^2} \right]$$

that the critical value of $|\rho|$ is that value which satisfies the equation

$$\frac{V(\bar{y}_{n_r})}{\sigma_y^2} = \frac{V(\bar{y}_p)}{\sigma_y^2} ,$$

whence

$$(29) \qquad \rho_c^2 = \frac{(C - nA)^2(n-3) + C(C - nA - nB)}{C(C - nA - nB)(n-2)}$$

where n is the optimum value of n at $\rho^2 = \rho_c^2$. The critical value of $|\rho|$ for the case under consideration is $|\rho_c| \sim .4152$, at which point $n \sim 36.944$; this may be verified by putting $n = 36.944$ in equation (29) to get

$$\rho_c^2 = \frac{[100 - 36.944(2.5)]^2 (36.944-3) + 100[100 - 36.944(2.5 + 0.1)]}{100[100 - 36.944(2.5 + 0.1)] (36.944-2)}$$

$$= .17231963 \qquad |\rho_c| = .4151$$

Intuitively, one would conjecture that for fixed A and B the critical value of $|\rho|$ would decrease were the total cost C allowed to increase; table 2 is in agreement with this assertion.

TABLE (2)

CRITICAL VALUES OF $|\rho|$ CORRESPONDING TO VALUES OF C = 2.5n + 0.1N

Cost C	Critical Value of $\|\rho\| = \|\rho_c\|$
100	.415
200	.405
300	.395
400	.393
500	.391
600	.390
700	.389
800	.389
900	.388
1000	.388
10000	.382

DOUBLE SAMPLING FROM EQUAL SIZED CLUSTERS

Cluster sampling is frequently more expedient in practice than simple random sampling; hence it was decided to extend the theory of double sampling to include the case where a random selection of clusters of population elements is subsampled at different rates for the cheap and the costly variates. In the initial stages an infinite population shall be considered with linear models of the form

$$X_{ij} = \bar{x}_p + b_i + w_{ij}$$

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{x}_p) + \varepsilon_{ij}$$

where X_{ij} = the value of X, the cheap variate, on the j'th element in i'th cluster

$\bar{x}_p$ = the population mean of X

$b_i = \bar{x}_{p_i} - \bar{x}_p$ = the deviation of the i'th cluster mean from the over all mean of X

$w_{ij} = X_{ij} - \bar{x}_{p_i}$ = the deviation of the ij'th element value from the i'th cluster mean. It is assumed that b_i and w_{ij} are normally and independently distributed with means of zero and variances of σ^2_{bx} and σ^2_{wx}, respectively.

Y_{ij} = the value of Y on the j'th element in the i'th cluster.

μ = the population mean of Y.

α_i = the vertical deviation of the i'th regression line from the average regression line over all clusters.

β = the population regression coefficient.

ε_{ij} = the deviation of Y_{ij} from the i'th regression line. It is assumed that α_i and ε_{ij} are independently distributed with means of zero and

variances of σ_a^2 and σ_ε^2, respectively.

The additional symbols which will be used are

$$\bar{y}_{p_i} = \mu + \alpha_i + \beta(\bar{x}_{p_i} - \bar{x}_p)$$

where $\bar{y}_{p_i}$ and $\bar{x}_{p_i}$ are the true means in the i'th cluster,

$$V(\bar{y}_{p_i}) = \sigma_{by}^2 = \sigma_a^2 + \beta^2\sigma_{bx}^2$$

$$\therefore \quad \sigma_a^2 = \sigma_{by}^2 - \beta^2\sigma_{bx}^2$$

and $$\sigma_{wy}^2 = E(Y_{ij} - \bar{y}_{p_i})^2 = \beta^2\sigma_{wx}^2 + \sigma_\varepsilon^2$$

$$\sigma_{wy}^2 = \rho^2\sigma_{wy}^2 + \sigma_\varepsilon^2$$

or $$\sigma_\varepsilon^2 = \sigma_{wy}^2(1-\rho^2) \quad .$$

Estimation

If we make a random selection of n clusters and draw random subsamples of size L on X and size m on Y, our estimate of μ is again of the form

$$\bar{y}_{ds} = \bar{y}_{nm} + b(\bar{x}_{nL} - \bar{x}_{nm}) \quad .$$

From the small sample of size nm we make our estimate of the regression equation:

$$\hat{Y}_{ij} = \bar{y}_{nm} + a_i + b(X_{ij} - \bar{x}_{nm}) \quad .$$

The least squares solutions for a_i and b are

$$\sum_j^m Y_{ij} - m\bar{y}_{nm} - ma_i - b\sum_j^m (X_{ij} - \bar{x}_{nm}) = 0$$

$$a_i = \bar{y}_{m_i} - \bar{y}_{nm} - b(\bar{x}_{m_i} - \bar{x}_{nm})$$

$$b = \frac{\sum\sum (X_{ij} - \bar{x}_{m_i})(Y_{ij} - \bar{y}_{m_i})}{\sum\sum (X_{ij} - \bar{x}_{m_i})^2} \qquad (30)$$

Thus

$$\hat{Y}_{ij} = \bar{y}_{nm} + \bar{y}_{m_i} - \bar{y}_{nm} - b(\bar{x}_{m_i} - \bar{x}_{nm}) + b(X_{ij} - \bar{x}_{m_i} + \bar{x}_{m_i} - \bar{x}_{nm})$$

$$(31) \qquad \hat{Y}_{ij} = \bar{y}_{m_i} + b(X_{ij} - \bar{x}_{m_i})$$

and the arithmetic mean of $\hat{Y}_{ij}$ over the entire sample is

$$\bar{y}_{ds} = \sum_{i=1}^{m} \sum_{j=1}^{L} \frac{\hat{Y}_{ij}}{nL} = \bar{y}_{nm} + b(\bar{x}_{nL} - \bar{x}_{nm}) \quad ,$$

which is an unbiased estimate of μ:

$$E(\bar{y}_{ds}) = E\left\{\mu + \bar{\alpha}_n + \beta(\bar{x}_{nm} - \bar{x}_p) + \bar{\varepsilon}_{nm} + (\bar{x}_{nL} - \bar{x}_{nm})\left[\beta + \frac{\Sigma\Sigma\varepsilon_{ij}(X - \bar{x}_i)}{\Sigma\Sigma(X - \bar{x}_i)^2}\right]\right\} = \mu \quad .$$

The variance of $\bar{y}_{ds}$ for the fixed sets of X selected in the sample is

$$(32) \qquad E(\bar{y}_{ds} - \mu)^2 = \frac{\sigma_\alpha^2}{n} + \sigma_\varepsilon^2\left[\frac{1}{nm} + \frac{(\bar{x}_{nL} - \bar{x}_{nm})^2}{\Sigma\Sigma(X_{ij} - \bar{x}_{m_i})^2}\right] + \beta^2(\bar{x}_{nL} - \bar{x}_p)^2 \quad .$$

The average value of this quantity over the entire range of X is

$$(33) \qquad V(\bar{y}_{ds}) = \frac{\sigma_\alpha^2}{n} + \sigma_\varepsilon^2\left[\frac{1}{nm} + \left(\frac{1}{nm} - \frac{1}{nL}\right)\left(\frac{1}{n(m-1)-2}\right)\right] + \beta^2\left(\frac{\sigma_{bx}^2}{n} + \frac{\sigma_{wx}^2}{nL}\right) \quad .$$

The method of estimating $V(\bar{y}_{ds})$ is not immediately apparent from (33). The variance formula may be simplified to a more familiar form:

$$(34) \qquad V(\bar{y}_{ds}) = \frac{\sigma_{by}^2}{n} + \sigma_{wy}^2(1 - \rho^2)\left[\frac{1}{nm} + \left(\frac{L-m}{nmL}\right)\left(\frac{1}{n(m-1)-2}\right)\right] + \frac{\rho^2\sigma_{wy}^2}{nL}$$

$$(35) \qquad V(\bar{y}_{ds}) = \frac{\sigma_{by}^2}{n} + \frac{\sigma_{wy}^2}{nm} - \frac{\sigma_{wy}^2}{nm}\left(\frac{L-m}{L}\right)\left(\frac{\rho^2[n(m-1)-1]-1}{n(m-1)-2}\right)$$

whereby it is seen that the third member on the right in (35) represents the reduction in error variance due to double sampling. An examination of (34) leads to a straight-forward method of estimation of variance since

$$s^2_{wy} = \frac{\Sigma\Sigma(Y_{ij} - \bar{y}_{m_i})^2}{n(m-1)}$$

$$E(s^2_{wy}) = \beta^2\sigma^2_{wx} + \sigma^2_{\varepsilon} = \sigma^2_{wy}$$

$$s^2_{by} = \frac{\Sigma(\bar{y}_{m_i} - \bar{y}_{nm})^2}{n-1}$$

$$E(s^2_{by}) = \sigma^2_{a} + \beta^2\sigma^2_{bx} + \frac{\beta^2\sigma^2_{wx}}{m} + \frac{\sigma^2_{\varepsilon}}{m} = \sigma^2_{by} + \frac{\sigma^2_{wy}}{m}$$

$$\frac{\Sigma\Sigma(Y_{ij} - \hat{Y}_i)^2}{n(m-1)-1} = \frac{(1-r^2)\Sigma\Sigma(Y_{ij} - \bar{y}_{m_i})^2}{n(m-1)-1} = \frac{s^2_{wy}\, n(m-1)(1-r^2)}{n(m-1)-1}$$

$$\frac{E[\,\Sigma\Sigma(Y_{ij} - \hat{Y}_i)^2\,]}{n(m-1)-1} = \sigma^2_{\varepsilon} = \sigma^2_{wy}(1-\rho^2)$$

then

$\frac{s^2_{by}}{n} - \frac{s^2_{wy}}{nm}$ is an unbiased estimate of $\frac{\sigma^2_{by}}{n}$

$\frac{s^2_{wy}\, n(m-1)(1-r^2)}{n(m-1)-1}$ is an unbiased estimate of $\sigma^2_{wy}(1-\rho^2)$

$\frac{s^2_{wy}[n(m-1)r^2-1]}{n(m-1)-1}$ is an unbiased estimate of $\rho^2\sigma^2_{wy}$

$\frac{(\bar{x}_{nL} - \bar{x}_{nm})^2}{\Sigma\Sigma(X_{ij} - \bar{x}_{m_i})^2}$ is an unbiased estimate of $E\left[\frac{(\bar{x}_{nL} - \bar{x}_{nm})^2}{\Sigma\Sigma(X_{ij} - \bar{x}_{m_i})^2}\right] = \frac{L-m}{nmL[n(m-1)-2]}$

$$(36)\quad \hat{V}(\bar{y}_{ds}) = \frac{s^2_{by}}{n} - \frac{s^2_{wy}}{nm}\left(\frac{L-m}{L}\right)\left(\frac{n(m-1)r^2-1}{n(m-1)-1}\right) + \frac{s^2_{wy}\, n(m-1)(1-r^2)(\bar{x}_{nL} - \bar{x}_{nm})^2}{[n(m-1)-1]\,\Sigma\Sigma(X_{ij} - \bar{x}_{m_i})^2}$$

If the finite population correction is introduced, (36) becomes

$$(37)\quad \hat{V}(\bar{y}_{ds}) = \frac{s^2_{by}}{n}\left(\frac{N-n}{N}\right) - \frac{s^2_{wy}}{nm}\left(\frac{L-m}{L}\right)\left(\frac{n(m-1)r^2-1}{n(m-1)-1}\right) + \frac{s^2_{wy}\,n(m-1)(1-r^2)(\bar{x}_{nL}-\bar{x}_{nm})^2}{[n(m-1)-1]\,\Sigma\Sigma(X_{ij}-\bar{x}_{m_i})^2}\ *$$

In practice the quantity $\frac{n(m-1)}{n(m-1)-1}$ will be sufficiently near unity that it may be ignored without introducing any noticeable bias. (36) then becomes

$$(38)\quad \hat{V}(\bar{y}_{ds}) = \frac{s^2_{by}}{n} - \frac{r^2 s^2_{wy}}{nm}\left(\frac{L-m}{L}\right) + s^2_{wy}(1-r^2)\left[\frac{(\bar{x}_{nL}-\bar{x}_{nm})^2}{\Sigma\Sigma(X_{ij}-\bar{x}_{m_i})^2}\right]$$

and (37) becomes

$$(39)\quad \hat{V}(\bar{y}_{ds}) \cong \frac{s^2_{by}}{n}\left(\frac{N-n}{N}\right) - \frac{r^2 s^2_{wy}}{nm}\left(\frac{L-m}{L}\right) + s^2_{wy}(1-r^2)\left[\frac{(\bar{x}_{nL}-\bar{x}_{nm})^2}{\Sigma\Sigma(X_{ij}-\bar{x}_{m_i})^2}\right] .$$

Two prominent features of this method of estimation now arise; (i) equation (36) is unbiased irrespective of the form of the distribution of the independent variable, (ii) the computations on the small sample on X and Y are identical to those of the simple analysis of covariance and may be carried through by the convenient procedure outlined in Chapter 12 of Snedecor's "Statistical Methods."

The restriction that clusters be of equal size is frequently difficult to satisfy in practice, as, for example, where one is sampling city blocks as clusters of urban households or square mile sections of land as clusters of rural households. The preceding theory does, however, lend itself to a combination of random and systematic sampling suggested by P. J. McCarthy and

---

* The reader may well question the absence of a correction term for the quantity $\frac{(\bar{x}_{nL}-\bar{x}_{nm})^2}{\Sigma\Sigma(X_{ij}-\bar{x}_{m_i})^2}$; this problem remains to be investigated.

others. If it is decided that the sampling rate for the X-variate shall be $\frac{1}{k} = \frac{n}{nk}$, it is seen that a systematic sample of every nk'th element may be looked upon as one completely enumerated cluster, and $\underline{n}$ such systematic samples selected by choosing $\underline{n}$ random starting point from the interval 1, 2, ..., nk constitute a sample of clusters which satisfies the requirements of randomness. Without knowledge of the total number of elements in the population it would be impossible, of course, to predetermine the necessarily random within cluster samples for the Y-variable; it would be necessary, in other words, to sample the clusters for the Y variable after the enumeration on X had been completed. If $\frac{1}{k} < .05$, $\frac{\underline{m}}{L} < .05$, where $\underline{m}$ is the within cluster sample size for Y and L is the size of the cluster, the f.p.c.'s might be ignored and the variance of the adjusted mean, $\bar{y}_{ds}$, of a sample of this type remains in the form

$$(40) \qquad V(\bar{y}_{ds}) = \frac{\sigma^2_{by}}{n} + \sigma^2_{wy}(1-\rho^2)\left[\frac{1}{nm} + E\,\frac{(\bar{x}_{nL} - \bar{x}_{nm})^2}{\Sigma\Sigma(X_{ij} - \bar{x}_{m_i})^2}\right] + \frac{\rho^2\sigma^2_{wy}}{nL} \quad .$$

SAMPLE DESIGN

The Curtis Impact Survey had three primary objectives: (1) to obtain the family characteristics, interests, possessions and buying habits of certain leading magazine households, (2) to obtain essential characteristics of the readers of the magazine and (3) to measure the impact of these magazines upon the readers.

The survey was divided into two double samples applying Use I and III mentioned above. Use I involved personal interviews with a national sample of 30,816 households from which certain basic data on magazine readers and family composition were obtained by an interview. This information was used for stratifying a sub-sample of 2,263 households from which the magazine readership and impact information was obtained. In two-thirds of the 2,263 households the number of readers within the households was obtained by asking the first person who came to the door how many readers there were in the household. This number is, of course, subject to error. Therefore, in one-third of the 2,263 households all adult persons in the household were interviewed to obtain actual readership. This subsample provided a basis for establishing the relationship between the two readership figures. This is an application of Use III in double sampling. The gain in efficiency is shown on page 32.

There are other features in the sample design than double sampling that bear discussion at this point. The sample design was a multi-stage probability sample based upon the "area" method in which the 833 sampling units averaged about 37 households per segment. The sampling was done in several stages. Primary sampling units consisting of counties or pseudo counties for the open country and cities for the urban population were stratified and sampled as follows.

The United States was divided into three primary strata, urban, village and open country. The urban strata consisted of all incorporated places having 2,500 or more people, the village strata consisted of all incorporated places having a population of less than 2,500 and all unincorporated places having a population of less than 2,500 and a population density of 100 or more people per square mile. The open country strata consisted of the rest of the population.

For the open country and village strata, the 3,056 counties were grouped into 100 substrata which were homogeneous in respect to livelihood, soil and agriculture type. One primary sampling unit was drawn from each of the 100 substrata with probability proportional to its size.

The urban strata were substratified by size and city and geographic location. Eight size classes were made from cities less than one million people. Within each of these size classes the cities were ordered by geographic areas averaging about one million people each. This resulted in 63 urban substrata from which one city was drawn with probability proportional to its size. Each of the primary units were then subsampled using segments as the secondary sampling unit.

ANALYSIS OF THE CURTIS IMPACT SAMPLE

A moot question facing the analyst of the sample described above is whether to regard households or individual household members as the units of observation. The character of major interest in this study is magazine readership and the choice of unit of observation will, in the main, determine the method of estimating total number of readers.

If the household is the unit, then, with regard to any particular magazine, each household may be characterized by two variables: X = number of readers as stated by a responsible adult, Y = number of actual readers according to the definition of reader used in the study. Reasonable methods of estimation are conceivable; in any event, however, the sampling error must necessarily be large since only a very small fraction of the households were interviewed completely to determine the number of actual readers. The computations required for such an estimate would tend to be prohibitive if one were to utilize the correlation between declared and actual readership in a statistically efficient manner.

If the individual household member is to be regarded as the unit of observation, then, again, each person is characterized by the two variables X and Y; in this instance, however, the number of declared or actual readers must be either 0 or 1. Thus the individual member is characterized by X = 1 or 0 according as the responsible adult member of his household declared him a reader or nonreader of the magazine, and Y = 1 or 0 according to whether he actually does or does not read the magazine. Intuitively, this concept is the more appealing of the two from the standpoint of both statistical and computational efficiency; statistical efficiency is enhanced in that information on nonreaders as well as readers may be fully utilized, and computations associated with samples from this type of population are in general relatively

simple. The apparent desirability of an analysis on this basis leads, then, to an investigation into the methods of estimation associated with double sampling from binomial-type populations.

Consider a sample segment drawn from the binomial population just described; the total number of household members in the segment is (say) $N_{..}$; there are (say) $N_{0.}$ individuals who would be declared nonreaders of magazine A by the responsible adults; and of these $N_{0.}$ individuals, N_{00} (say) actually do not read the magazine while N_{01} (say) actually do read it. With this notation the $N_{..}$ individuals may be classified into a two-way table:

Y \ X	0	1	
0	N_{00}	N_{10}	$N_{.0}$
1	N_{01}	N_{11}	$N_{.1}$
	$N_{0.}$	$N_{1.}$	$N_{..}$

The enumerator obtains a statement from a responsible adult member of each household in the segment and so determines the marginal totals $N_{0.}$ and $N_{1.}$. A sample of (say) $n_{..}$ individuals are interviewed to determine actual readership; thus both X and Y information is known on $n_{..}$ persons and they may be classified into a sample two-way table:

Y \ X	0	1	
0	n_{00}	n_{10}	$n_{.0}$
1	n_{01}	n_{11}	$n_{.1}$
	$n_{0.}$	$n_{1.}$	$n_{..}$

The problem, then, is to estimate the marginal total $N_{.1}$ given the sample table and the marginal totals $N_{0.}$ and $N_{1.}$. One logical estimate might be

$\hat{N}_{\cdot 1} = \frac{n_{\cdot 1}}{n_{\cdot\cdot}} N_{\cdot\cdot}$; i.e., the sample proportion of actual readers applied to the total number of persons in the segment. This estimate does not, of course, utilize the X information. The more reasonable estimate (and also the least squares solution) which does utilize the X information would be

$$\hat{N}_{\cdot 1} = \frac{n_{01}}{n_{0\cdot}} N_{0\cdot} + \frac{n_{11}}{n_{1\cdot}} N_{1\cdot}$$

where $\frac{n_{01}}{n_{0\cdot}}$ is the sample proportion of declared nonreaders who were actually readers and $\frac{n_{11}}{n_{1\cdot}}$ is the sample proportion of declared readers who were actually readers.

The sampling errors of these two estimates cannot be computed for the sample as it was actually taken in the field - first, because the sample is not repeatable since there was no predetermined procedure for choosing one of the responsible adult members of the household to act as a respondent and second, because the respondents subjected to the intensive interview were not selected by a simple random process. It is possible, however, to envisage a sampling scheme which is substantially equivalent to that followed in the field and which does allow formulation of sampling error. For example, the predetermined scheme for selecting the responsible adult respondent from a household might be to first determine which of the responsible adults is at home most frequently and then interview that person - or, perhaps, to always interview the female head, when one exists; either of these schemes would closely approximate the results actually obtained in the survey. A random selection of respondents for intensive interview was also approximated in the field since the enumerator did not purposely select these individuals. Assuming, then, that "responsible adult member" is a mutually exclusive property and that the sample of size $n_{\cdot\cdot}$ was drawn by a random process, the

variance of the second estimate is:

$$V\left[\frac{n_{01}}{n_{0\bullet}} N_{0\bullet} + \frac{n_{11}}{n_{1\bullet}} N_{1\bullet}\right] = \frac{\frac{N_{01}N_{00}}{N_{0\bullet}-1} \sum_{i=1}^{n_{\bullet\bullet}-1} \binom{N_{0\bullet}}{n_{\bullet\bullet}-i}\binom{N_{1\bullet}}{i} \frac{N_{0\bullet}-n_{\bullet\bullet}+i}{n_{\bullet\bullet}-i}}{\sum_{i=1}^{n_{\bullet\bullet}-1} \binom{N_{0\bullet}}{n_{\bullet\bullet}-i}\binom{N_{1\bullet}}{i}}$$

$$+ \frac{\frac{N_{10}N_{11}}{N_{1\bullet}-1} \sum_{i=1}^{n_{\bullet\bullet}-1} \binom{N_{1\bullet}}{n_{\bullet\bullet}-i}\binom{N_{0\bullet}}{i} \frac{N_{1\bullet}-n_{\bullet\bullet}+i}{n_{\bullet\bullet}-i}}{\sum_{i=1}^{n_{\bullet\bullet}-1} \binom{N_{0\bullet}}{n_{\bullet\bullet}-i}\binom{N_{1\bullet}}{i}}$$

and an unbiased estimate of this sampling error is:

$$\hat{V}\left[\frac{n_{01}}{n_{0\bullet}} N_{0\bullet} + \frac{n_{11}}{n_{1\bullet}} N_{1\bullet}\right] = \frac{N_{0\bullet}(N_{0\bullet}-n_{0\bullet})}{n_{0\bullet}-1} \frac{n_{00}n_{01}}{n_{0\bullet}^2} + \frac{N_{1\bullet}(N_{1\bullet}-n_{1\bullet})}{n_{1\bullet}-1} \frac{n_{10}n_{11}}{n_{1\bullet}^2}$$

The variance of the first estimate is:

$$V\left[\frac{n_{\bullet 1}}{n_{\bullet\bullet}} N_{\bullet\bullet}\right] = \frac{N_{\bullet 0}N_{\bullet 1}(N_{\bullet\bullet}-n_{\bullet\bullet})}{n_{\bullet\bullet}(N_{\bullet\bullet}-1}$$

and an unbiased estimate of this sampling error is:

$$\hat{V}\left[\frac{n_{\bullet 1}}{n_{\bullet\bullet}} N_{\bullet\bullet}\right] = \frac{N_{\bullet\bullet}(N_{\bullet\bullet}-n_{\bullet\bullet})}{n_{\bullet\bullet}-1} \frac{n_{\bullet 0}n_{\bullet 1}}{n_{\bullet\bullet}^2}$$

One would expect a gain in efficiency due to double sampling if the correlation between declared and actual readership were sufficiently high; the comparison of the two error variances, $V\left[\frac{n_{\bullet 1}}{n_{\bullet\bullet}} N_{\bullet\bullet}\right]$ and $V\left[\frac{n_{01}}{n_{0\bullet}} N_{0\bullet} + \frac{n_{11}}{n_{1\bullet}} N_{1\bullet}\right]$, is not, however, completely justified when costs are brought into consideration. The first estimate requires knowledge of only the total number of persons, $N_{\bullet\bullet}$, in the segment, along with the small sample measuring actual readership; the second estimate requires a statement of readership

from every household, along with a small sample measuring actual readership. For a fixed total cost of sampling the segment one could probably afford to measure more persons for actual readership in the first case since the cost of counting the number of household members can hardly be more than the cost of obtaining a readership statement from a responsible adult member of each household. One is lead, then, to the consideration of two distinct cost functions:

$$C = n_{\cdot\cdot} A + N_{\cdot\cdot} B \qquad\qquad C = k_{\cdot\cdot} A + N_{\cdot\cdot} D$$

where C = total cost of the sample

A = average cost of measuring actual readership of an individual

B = average cost per individual of obtaining the statements from the responsible adults.

D = average cost of counting an individual

$k_{\cdot\cdot}$ = number of individuals measured for actual readership in the case where no X-information is required.

The efficiency of the double-sample estimate relative to the single sample estimate may then be computed as

$$R.E. = \frac{V\left[\frac{k_{\cdot 1}}{k_{\cdot\cdot}} N_{\cdot\cdot}\right] \quad (100)}{V\left[\frac{n_{01}}{n_{0\cdot}} N_{0\cdot} + \frac{n_{11}}{n_{1\cdot}} N_{1\cdot}\right]}$$

This expression, however, is quite unwieldy algebraically because of the combinatorial factors in the denominator. Superficial examination of the ratio of these variances thus reveals very little information on their relative magnitude, and it will be convenient to consider a special case of the double sample design which does allow for easy interpretation of the relative efficiency.

Suppose, for example, that the enumerator were to stratify the segment population on the basis of the variable X; i.e., suppose that after obtaining readership information from each of the uniquely defined "responsible adults," the enumerator then randomly selected a fixed number $n_{0\bullet}$ (say) of the declared nonreaders and $n_{1\bullet}$ (say) of the declared readers for intensive interviewing. With this modification imposed upon the double sample design, the variance of the estimate simplifies to

$$V\left[\frac{n_{01}}{n_{0\bullet}} N_{0\bullet} + \frac{n_{11}}{n_{1\bullet}} N_{1\bullet}\right] = \frac{N_{01}N_{00}(N_{0\bullet}-n_{0\bullet})}{n_{0\bullet}(N_{0\bullet}-1)} + \frac{N_{11}N_{10}(N_{1\bullet}-n_{1\bullet})}{n_{1\bullet}(N_{1\bullet}-1)}$$

and, replacing $N_{0\bullet}-1$ and $N_{1\bullet}-1$ by $N_{0\bullet}$ and $N_{1\bullet}$, respectively,

$$V\left[\frac{n_{01}}{n_{0\bullet}} N_{0\bullet} + \frac{n_{11}}{n_{1\bullet}} N_{1\bullet}\right] \sim \frac{N_{01}N_{00}(N_{0\bullet}-n_{0\bullet})}{n_{0\bullet}N_{0\bullet}} + \frac{N_{11}N_{10}(N_{1\bullet}-n_{1\bullet})}{n_{1\bullet}N_{1\bullet}};$$

likewise,

$$V\left[\frac{k_{\bullet 1}}{k_{\bullet\bullet}} N_{\bullet\bullet}\right] \sim \frac{N_{\bullet 0}N_{\bullet 1}(N_{\bullet\bullet}-k_{\bullet\bullet})}{k_{\bullet\bullet}N_{\bullet\bullet}}$$

and

$$R.E. \sim \frac{\dfrac{N_{\bullet 0}N_{\bullet 1}(N_{\bullet\bullet}-k_{\bullet\bullet})}{k_{\bullet\bullet}N_{\bullet\bullet}}}{\dfrac{N_{01}N_{00}(N_{0\bullet}-n_{0\bullet})}{N_{0\bullet}n_{0\bullet}} + \dfrac{N_{10}N_{11}(N_{1\bullet}-n_{1\bullet})}{n_{1\bullet}N_{1\bullet}}} \quad (100)$$

As a further simplification it may be supposed that $n_{0\bullet}$ and $n_{1\bullet}$ are chosen proportional to $N_{0\bullet}$ and $N_{1\bullet}$, respectively; i.e., $n_{0\bullet} = r\,N_{0\bullet}$ and $n_{1\bullet} = r\,N_{1\bullet}$ where $r = \dfrac{n_{\bullet\bullet}}{N_{\bullet\bullet}}$. It may be shown, incidentally, that proportional allocation is not the optimum allocation in this case; in fact, the optimum value of $n_{0\bullet}$ is

$$n_{0\bullet} \sim n_{\bullet\bullet}\left[\frac{N_{01}N_{00} - \sqrt{N_{01}N_{00}N_{11}N_{10}}}{N_{01}N_{00} - N_{11}N_{10}}\right].$$

Using proportional allocation, however, the relative efficiency becomes simply

$$R.E. \sim \frac{r(1-t)}{t(1-r)} \quad \frac{N_{\cdot 0}N_{\cdot 1}N_{1\cdot}N_{0\cdot}}{N_{\cdot 0}N_{\cdot 1}N_{1\cdot}N_{0\cdot}-(N_{11}N_{00}-N_{01}N_{10})^2}$$

$$\text{where} \quad t = \frac{k_{\cdot\cdot}}{N_{\cdot\cdot}}$$

which, in turn, reduces to

$$R.E. \sim \frac{r(1-t)}{t(1-r)} \quad \frac{1}{1-\rho^2}$$

when ρ is defined to be

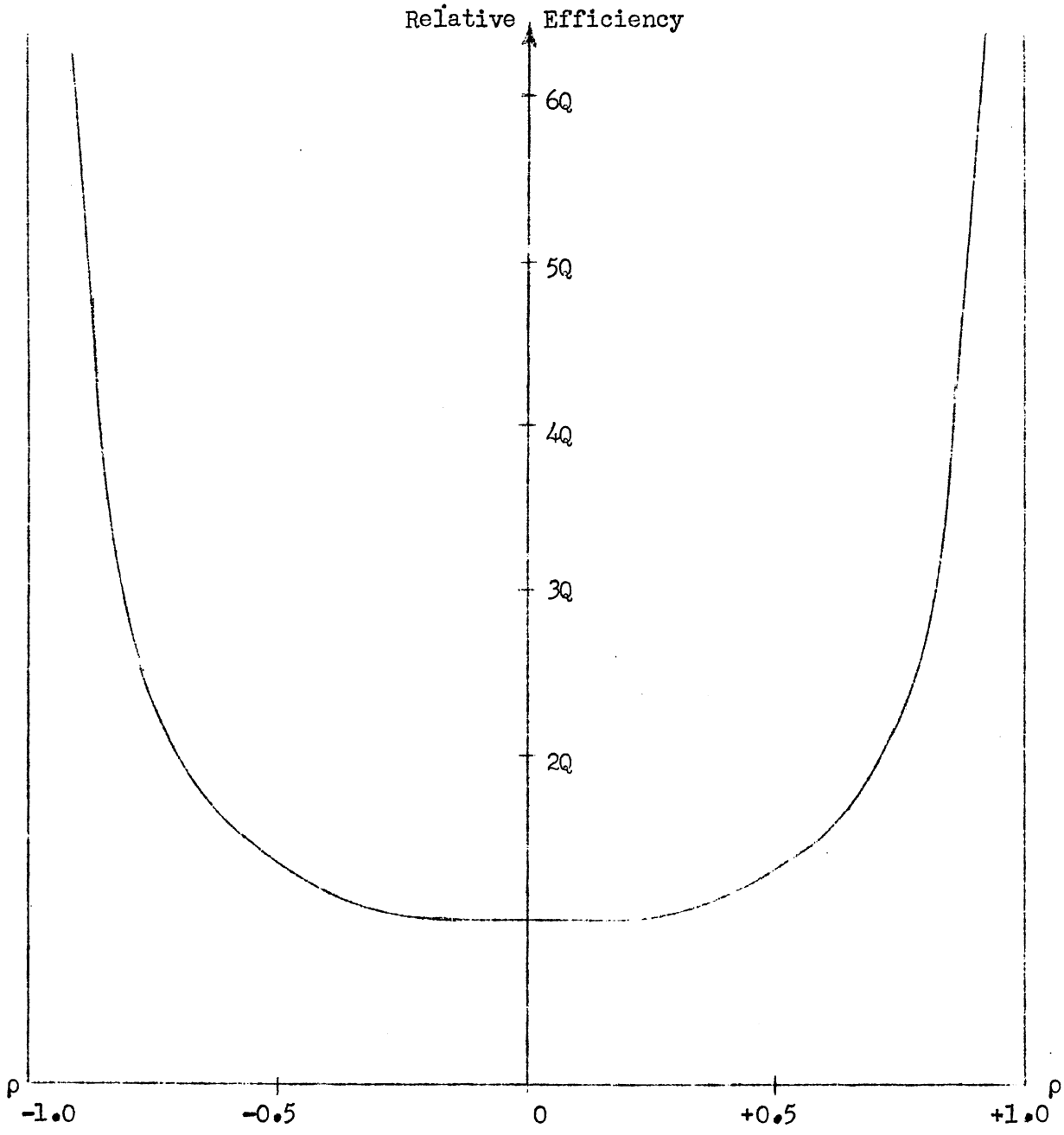
$$\rho = \frac{\sum_{i=1}^{N_{\cdot\cdot}} (X_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum_{1}^{N_{\cdot\cdot}} (X_i - \bar{x})^2 \sum_{1}^{N_{\cdot\cdot}} (Y_i - \bar{y})^2}} = \frac{N_{11}N_{00}-N_{01}N_{10}}{\sqrt{N_{\cdot 1}N_{1\cdot}N_{\cdot 0}N_{0\cdot}}}$$

In this simple form the relative efficiency of double sampling over single sampling is seen to depend entirely upon the correlation between the readership statements of the responsible adults and the actual readership of the household members (assuming fixed costs). The graph of the function $\varphi(\rho) = \frac{Q}{1-\rho^2}$ reveals the tremendous potential gains inherent in the double sampling technique as applied to binomial type populations [Figure (2)].

The discussion to this point has been centered upon the final sampling stage of the Curtis Impact Survey; in fact, the national sample was multistage, as described earlier, but for the purpose of determining whether double sampling allows for more efficient estimation than single sampling one may ignore the stages leading up to the selection of sample segments. The method of estimation in either case involves first the estimation of the segment totals and then a common procedure of combining and expanding to a national estimate. It is quite possible that the other components of sampling error, viz., among segments

FIGURE (2)

THE GRAPH OF R.E. = $\frac{Q}{1-\rho^2}$



within primary sampling units and among primary sampling units within strata, may completely overshadow the variance within segments; this fact, however has no bearing upon the question at hand - that is, did double sampling actually allow a more efficient estimate in this survey than would single sampling?

Another point which has thus far been disregarded is the possibility of a "Don't Know" response from the responsible adult member of a household. The introduction of this additional category would expand the two-way table to

		X			
		Nonreader	Don't Know	Reader	
Y	Non reader	N_{00}	N_{10}	N_{20}	$N_{\cdot 0}$
	Reader	N_{01}	N_{11}	N_{21}	$N_{\cdot 1}$
		$N_{0\cdot}$	$N_{1\cdot}$	$N_{2\cdot}$	$N_{\cdot\cdot}$

and the sample table would be obtained by replacing the N_{ij} by n_{ij}.

The estimate is then

$$\hat{N}_{\cdot 1} = \frac{n_{01}}{n_{0\cdot}} N_{0\cdot} + \frac{n_{11}}{n_{1\cdot}} N_{1\cdot} + \frac{n_{21}}{n_{2\cdot}} N_{2\cdot}$$

with variance

$$V = \frac{N_{00}N_{01}(N_{0\cdot}-n_{0\cdot})}{n_{0\cdot}(N_{0\cdot}-1)} + \frac{N_{10}N_{11}(N_{1\cdot}-n_{1\cdot})}{n_{1\cdot}(N_{1\cdot}-1)} + \frac{N_{20}N_{21}(N_{2\cdot}-n_{2\cdot})}{n_{2\cdot}(N_{2\cdot}-1)}$$

in the case where the segment is stratified on the basis of X, and a more lengthy expression in the case where the segment is not stratified. (This latter expression corresponds to that given for the two-way cast but is of little interest in this discussion). In either case, the unbiased estimate of variance is

$$\hat{V} = \frac{N_{0\cdot}(N_{0\cdot}-n_{0\cdot})}{n_{0\cdot}-1} \frac{n_{00}n_{01}}{n_{0\cdot}^2} + \frac{N_{1\cdot}(N_{1\cdot}-n_{1\cdot})}{n_{1\cdot}-1} \frac{n_{10}n_{11}}{n_{1\cdot}^2} + \frac{N_{2\cdot}(N_{2\cdot}-n_{2\cdot})}{n_{2\cdot}-1} \frac{n_{20}n_{21}}{n_{2\cdot}^2}$$

The reason for omitting this third category from the general discussion is that it raises other problems which do not resolve into such simple expressions as

the function which is plotted in Figure (2). In obtaining numerical estimates from the sample data it was necessary, of course, to take into account the "Don't Know" responses.

To date, data from the sample segments of the rural place and open country zones have been analyzed to obtain estimates of segment totals and the corresponding sampling errors. Not all of the within segment samples on actual readership were sufficiently large to permit estimation by the forementioned procedure, i.e., in some segments, for example, none of the declared nonreaders were sampled to determine actual readership - in which case $n_{2\cdot} = 0$ and $\hat{N}_{\cdot 1}$ is undefined. Those segments which were sampled sufficiently tended to be the larger ones and hence the average size of segment in this study is larger than the overall sample average.

The numerical results, calculated on an average per segment basis, are summarized in the following table

	Magazine A				Magazine B				Magazine C				Magazine D				Average			
$\hat{V}(\hat{N}_{\cdot 1})$	110.3				68.34				37.23				43.94				64.95			
Estimated Cell Frequencies	66	0	4	70	72	1	5	78	73	1	4	78	81	1	2	84	73	1	4	78
	5	1	7	13	2	0	14	16	1	0	6	7	1	0	7	8	2	0	8	10
	71	1	11	83	74	1	19	94	74	1	10	85	82	1	9	92	75	1	12	88
$\hat{\rho}$ (ignoring DK's)	.59				.76				.69				.80				.70			

Estimated cell frequencies were obtained from a larger number of sample segments than were available for estimates of variance since the latter require that $n_{0\cdot}, n_{1\cdot}, n_{2\cdot} > 1$ accordingly as $N_{0\cdot}, N_{1\cdot}, N_{2\cdot} > 1$. The average values of $n_{0\cdot}, n_{1\cdot}, n_{2\cdot}$ from the segments for which estimates of cell frequencies were obtained are $\bar{n}_{0\cdot} \sim 3.0$, $\bar{n}_{1\cdot} \sim 0$, $\bar{n}_{2\cdot} \sim 1.3$ (averaged over all magazines).

As an approach to the estimation of the efficiency of double sampling relative to single sampling , consider a segment population distributed like the

average in the above table and, for simplicity, ignore the "Don't Know" category, i.e.,

$N_{00} = 73$	$N_{10} = 4$	$N_{\cdot 0} = 77$
$N_{01} = 2$	$N_{11} = 8$	$N_{\cdot 1} = 10$
$N_{0\cdot} = 75$	$N_{1\cdot} = 12$	$N_{\cdot\cdot} = 87$

Then for samples of size $n_{0\cdot} = 3.0$ and $n_{1\cdot} = 1.3$ the variance of the estimate is

$$V\left[\frac{n_{01}}{n_{0\cdot}} N_{0\cdot} + \frac{n_{11}}{n_{1\cdot}} N_{1\cdot}\right] = \frac{73(2)(75-3)}{3(75-1)} + \frac{4(8)(12-1.3)}{1.3(12-1)} = 71.29$$

while the variance of a single sample estimate from a sample of size $3.0+1.3=4.3$ is

$$V\left[\frac{n_{\cdot 1}}{n_{\cdot\cdot}} N_{\cdot\cdot}\right] = \frac{77(10)(87-4.3)}{4.3(87-1)} = 172.20$$

and the relative efficiency is

$$R.E. = \frac{172.20}{71.29}(100) = 241.5\% .$$

If the cost functions described earlier are considered, i.e.,

$$C = n_{\cdot\cdot}A + N_{\cdot\cdot}B = k_{\cdot\cdot}A + N_{\cdot\cdot}D$$

then the size of the single sample, $k_{\cdot\cdot}$, corresponding to $n_{\cdot\cdot} = 4.3$, $N_{\cdot\cdot} = 87$, is

$$k_{\cdot\cdot} = 4.3 + 87\left(\frac{B-D}{A}\right) .$$

In this study, the cost A was approximately 10 times as large as B and so

$$k_{\cdot\cdot} = 13 - 8.7\left(\frac{D}{B}\right)$$

Now the value of $k_{\cdot\cdot}$ for which $V\left[\frac{k_{\cdot 1}}{k_{\cdot\cdot}} N_{\cdot\cdot}\right] = V\left[\frac{n_{01}}{n_{0\cdot}} N_{0\cdot} + \frac{n_{11}}{n_{1\cdot}} N_{1\cdot}\right] = 71.29$ is $k_{\cdot\cdot} = 9.7$; hence if the ratio $\left(\frac{D}{B}\right)$ is greater than 0.38 one would conclude that double sampling offered real gains with respect to the estimation of magazine readership.

DISCUSSION

I. LIMITATIONS IN THE THEORY

The estimation procedure outlined above for double sampling from a population distributed in a 2x2 table may be easily extended to include populations classified into any rxc table. The estimates of totals and estimates of sampling error are independent of the numerical values assigned to the several levels of the X and Y variates. Correlation, on the other hand, is dependent upon the relative weights placed upon each row and column in the table, and in practice these weights are frequently unknown. An example is given in this study when the "Don't Know" category is introduced - if the numerical equivalent to "no" or "nonreader" is 0 and to "yes" or "reader" is 1, then what is the numerical equivalent to "Don't Know"? It is to be emphasized, however, that the lack of a satisfactory measure of correlation does in no way affect the desirability of the double sample estimates.

An assumption of normality was introduced in the investigation of optimum allocation of resources to the large and small samples for a regression estimate of the population mean. This assumption is rarely fulfilled in practice, particularly in studies of economic or social characteristics of human populations. It is therefore important to note that normality was invoked merely to facilitate the operation of E(expectation) on the quotient $\dfrac{\frac{(\bar{x}_L - \bar{x}_n)^2}{n}}{\sum_{i=1}^{} (X_i - \bar{x}_n)^2}$ and that even without normality $E\left[\dfrac{(\bar{x}_L - \bar{x}_n)^2}{\Sigma(X_i - \bar{x}_n)^2}\right]$ is logically of negligible size.

The approximation for n which proved satisfactory in the case of a normally distributed X-variable may thus be assumed satisfactory in the more general case.

II. APPLICATIONS OF DOUBLE SAMPLING

Examples of the practical application of the double sampling technique in its several forms are numerous both in the literature and in unpublished works. A brief list of publications describing applications of double sampling is attached and it will suffice here to mention but a few additional situations in which the technique has been or could be applied.

The National Analysts, Inc., have on several occasions found use for Neyman's scheme of stratification on the basis of the cheap variable; an example of this was a survey of retail outlets for household appliances. The cheap variable X in this case was an approximate measure of size of sales; the large sample of outlets drawn in the first stage was then stratified on the basis of sales and the strata were subsequently sampled proportionate to size for measurement of the more costly variables.

It is also an opinion held in that organization that the trend in commercial sampling will be toward fuller use of the double sampling technique. After the fashion of Jessen's matched sampling, a survey conducted at one point in time may be utilized in connection with a subsequent survey to enhance the precision of the second estimate. For example, if the first sample includes m segments from a given primary sampling unit and the second sample is selected to contain r different segments from the same PSU and if, as in any well-conducted survey, certain control variables were measured on the first sample, then the appropriate choice of a control common to both samples can lead to estimates with efficiencies comparable to that pictured graphically in the text.

A few other examples in which double sampling has been used are (1) growth studies on plants and animals where one measurement, such as length, is easily obtained while another, such as circumference or depth, requires greater care; (2) quantitative chemical analyses which can be performed crudely by some simple procedure and precisely by some other time-consuming procedure; and (3) more generally, situations in which an eye estimate may be compared to a precise measurement.

LITERATURE ON DOUBLE SAMPLING

*1. Bose, Chameli. Note on sampling error in the method of double sampling. Sankhya 6:329-330, 1943

2. Cochran, W. G. Sampling theory when the units are of unequal sizes. Jour. Amer. Stat. Assoc. 37:199-212, 1942

*3. _____. Sample survey techniques. Mimeo. notes, 1948

*4. Jessen, R. J. Statistical investigation of a sample survey for obtaining farm facts. Iowa Agri. Exp. Sta. Res. Bul. 304, 1942

*5. Neyman, J. Contribution to the theory of sampling human populations. Jour. Amer. Stat. Assoc. 33:101-116, 1938.

*6. Schumacher, F. X., and Chapman, R. A. Sampling methods in forestry and range management. Duke Univ. School of For. Bul. 7, 1942

7. Snedecor, G. W., and King, A. J. Recent developments in sampling for agricultural statistics. Jour. Amer. Stat. Assoc. 37:95-102, 1942

8. Watson, D. J. The estimation of leaf areas. Jour. Agri. Sci. 27:474, 1937

9. Wilm, H. G., Costello, D. F., and Klipple, G. E. Estimating for age yield by the double sampling method. Jour. Amer. Soc. Agron. 36:194-203, 1944

10. Yates, F. Sampling for surveys and censuses, Griffin, 1948

*cited

Appendix I

Notation:

N = size of the large random sample on X

N_i = size of the i'th group, i.e., $\sum_{i=1}^{k} N_i = N$

n_i = size of the small random sample on Y, drawn from the i'th group

$w_i = \frac{N_i}{N}$ = estimated proportion of population elements falling in the i'th stratum

W_i = true proportion of population elements falling in the i'th stratum

μ = population mean of the variable Y

μ_i = true mean of the i'th stratum, i.e., $\mu = \sum_{i=1}^{k} W_i \mu_i$

$\bar{y}_i = \frac{\sum_{j}^{n_i} Y_{ij}}{n_i} = \hat{\mu}_i$ = estimate of the i'th stratum mean

$\bar{y}_p = \Sigma w_i \bar{y}_i = \hat{\mu}$ = estimate of the population mean

Assertion:

$$V(\bar{y}_p) = \sum_{i=1}^{k} \left[\frac{\sigma_i^2}{n_i} \left\{ W_i^2 + \frac{W_i(1-W_i)}{N} \right\} + \frac{W_i(\mu_i-\mu)^2}{N} \right]$$

Proof:

Write $w_i = W_i + \alpha_i$

$\bar{y}_i = \mu_i + \varepsilon_i$

Then $\bar{y}_p - \mu = \sum_{i}^{k} (w_i \bar{y}_i - W_i \mu_i) = \sum_{i}^{k} (W_i \varepsilon_i + \mu_i \alpha_i + \alpha_i \varepsilon_i)$

$V(\bar{y}_p) = E(\bar{y}_p - \mu)^2 = E\left[\sum_{i}^{k} (W_i \varepsilon_i + \mu_i \alpha_i + \alpha_i \varepsilon_i) \right]^2$

where $E(\varepsilon_i^2) = \frac{\sigma_i^2}{n_i}$

$E(\alpha_i^2) = \frac{W_i(1-W_i)}{N}$

$$E(\varepsilon_i \varepsilon_j) = 0 \qquad i \neq j$$

$$E(a_i a_j) = -\frac{W_i W_j}{N} \qquad i \neq j$$

$$\therefore \quad V(\bar{y}_p) = \sum_i^k \frac{W_i^2 \sigma_i^2}{n_i} + \sum_i^k \frac{W_i(\mu_i - \mu)^2}{N} + \sum_i^k \frac{W_i(1 - W_i)}{N} \frac{\sigma_i^2}{n_i}$$

Appendix II

Notation:

$$V(\bar{y}_p) = \sum_{i=1}^{k} \left[\frac{\sigma_i^2}{n_i} \left\{ W_i^2 + \frac{W_i(1-W_i)}{N} \right\} + \frac{W_i(\mu_i - \mu)^2}{N} \right]$$

where the symbols are as given in Appendix I

$C = nA + NB$

where C = total cost of sample

A = cost of measuring the variable Y on a sample element

B = cost of measuring the variable X on a sample element

Assertion:

$V(\bar{y}_p)$ is approximately minimized for $\frac{n}{N} = \sum_{i}^{k} W_i \sigma_i \sqrt{\frac{B}{A \Sigma W_i(\mu_i - \mu)}}$

when C is fixed.

Proof:

Write $F = V(\bar{y}_p) + \lambda(C - nA - NB)$

$$\frac{\delta F}{\delta n_i} = -\frac{\sigma_i^2}{n_i^2} \left[W_i^2 + \frac{W_i(1-W_i)}{L} \right] = \lambda A$$

$\therefore$ n_i is proportional to $\sigma_i \sqrt{W_i^2 + \frac{W_i(1-W_i)}{N}}$, but if the second term in the radical is considered negligible we may write

$$(1) \qquad n_i = \frac{nW_i\sigma_i}{\sum_i W_i\sigma_i}$$

Write $F_1 = V_1(\bar{y}_p) + \lambda(C - nA - NB)$

where

$$V_1 = \frac{(\Sigma W_i\sigma_i)^2}{n} + \frac{\Sigma W_i(\mu_i - \mu)^2}{N}$$

, replacing n_i by (1) in V and dropping $\frac{W_i(1-W_i)}{N}$

then set $\frac{\delta F_1}{\delta n}$, $\frac{\delta F_1}{\delta N} = 0$

Appendix III

Notation: See text

Assertion:

$$V(\bar{y}_p) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_i (X_i - \bar{x}_n)^2} \right] + \beta^2 (\bar{x}_N - \mu_x)^2 \qquad \text{for fixed X's}$$

Proof:

$$V(\bar{y}_p) = E(\bar{y}_p - \mu)^2$$

$$= E\left[\bar{\varepsilon}_n + \beta(\bar{x}_N - \mu_x) + (\bar{x}_N - \bar{x}_n) \frac{\Sigma \varepsilon (X - \bar{x}_n)}{\Sigma (X - \bar{x}_n)^2} \right]^2$$

$$= E(\bar{\varepsilon}_n)^2 + \beta^2 (\bar{x}_N - \mu_x)^2 + (\bar{x}_N - \bar{x}_n)^2 \, E\left[\frac{\Sigma \varepsilon (X - \bar{x}_n)}{\Sigma (X - \bar{x}_n)^2} \right]^2$$

$$+ 2\beta(\bar{x}_N - \mu_x)(\bar{x}_N - \bar{x}_n) \; E\left[\frac{\Sigma \varepsilon (X - \bar{x}_n)}{\Sigma (X - \bar{x}_n)^2} \right] + 2E\left[\bar{\varepsilon}_n \frac{\Sigma \varepsilon (X - \bar{x}_n)}{\Sigma (X - \bar{x}_n)^2} \right]$$

$$= \frac{\sigma_\varepsilon^2}{n} + \beta^2 (\bar{x}_N - \mu_x)^2 + \sigma_\varepsilon^2 \frac{(\bar{x}_N - \bar{x}_n)^2}{\Sigma (X - \bar{x}_n)^2} + 0 + 0$$

Appendix IV

Notation: See text

Assertion:

$$E_x\left[V(\bar{y}_p)\right] = \sigma_\varepsilon^2\left[\frac{1}{n} + \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{1}{n-3}\right)\right] + \frac{\beta^2\sigma_x^2}{N}$$

when E_x means E operating on X and where $V(\bar{y}_p)$ is as given in Appendix III

Proof:

$$1^{\circ}\quad E\left[\beta^2(\bar{x}_N - \mu_x)^2\right] = \frac{\beta^2\sigma_x^2}{N}$$

$$2^{\circ}\quad \frac{(\bar{x}_N - \bar{x}_n)^2}{\Sigma\,(X - \bar{x}_n)^2} \text{ is distributed like } \frac{N-n}{Nn(n-1)}\,F_{1,n-1}$$

where F denotes Snedecor's F. This may be proven by a simple orthogonal transformation.

$$\therefore\quad E\left[\frac{(\bar{x}_N - \bar{x}_n)^2}{\Sigma\,(X - \bar{x}_n)^2}\right] = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{1}{n-3}\right) \quad \text{since } E(F) = \frac{n-1}{n-3}$$

Notation: See text

Assertion:

$$V(\bar{y}_w) = \frac{V(\bar{y}_u)\, V(\bar{y}_p)}{V(\bar{y}_u)+V(\bar{y}_p)} \text{ is approximately minimized at } \frac{n}{N} = \frac{\sqrt{1-\rho^2} - (1-\rho^2)}{\rho^2}$$

when N is fixed and m + n is fixed at C.

Proof:

Write $F(n,m) = f(n,m) + \lambda(m+n-C)$

where

$$f(n,m) = \frac{V(\bar{y}_u)\, g(n)}{V(\bar{y}_u)+g(n)}$$

where

$$g(n) = \frac{\sigma_y^2(1-\rho^2)}{n} + \frac{\sigma_y^2\,\rho^2}{N} \sim V(\bar{y}_p) .$$

Then

$$\frac{\delta F}{\delta n} : \frac{\delta f}{\delta n} = \lambda$$

$$\frac{\delta F}{\delta m} : \frac{\delta f}{\delta m} = \lambda$$

$$\therefore \quad \frac{\delta f}{\delta n} = \frac{\delta f}{\delta m}$$

$$\therefore \quad [V(\bar{y}_u)]^2 \frac{\delta g}{\delta n} = [g(n)]^2 \frac{\delta [V(\bar{y}_u)]}{\delta m}$$

$$\frac{(1-\rho^2)}{n^2} = \frac{(1-\rho^2)^2}{n^2} + \frac{2\rho^2(1-\rho^2)}{nN} + \frac{\rho^4}{N^2}$$

$$\therefore \quad n = \frac{\sqrt{1-\rho^2} - (1-\rho^2)}{\rho^2 / N}$$